Play Store App Review Analysis

Saniya Bubere,
Rahul Jadhav,
Ashi Saxena,
Rishabh Rohil.
Data science trainees,
Alma Better, Bangalore

Abstract:

We are analyzing the data of playstore apps and user reviews .Our objective is to Actionable insights that can be drawn for developers to work on and capture the Android market and key factors responsible for app engagement and success.

We studied two data sets one in terms of app category , its size , no of installed , no of reviews ,its price , for whom these apps can suitable , genres app belong to , app last update, current version , for which android version it is suitable and another dataset is of user reviews in terms of user review , their sentiment ,sentiment polarity and sentiment subjectivity.

***Key Words*:** Google Play Store Apps, Ratings Prediction, Exploratory Data Analysis, Machine Learning.

1. Problem Statement

Data is taken from the Google play store dataset. Every row contains various entries regarding a certain app. We will be doing Exploratory data analysis on this data set, which is a very important step in the data science cycle, as it not only helps in taking very initial business decisions .

Objectives:

- Exploring and Cleaning the Dataset
- To establish relationship between various features of the Dataset
- Present these relationships using Various data visualization techniques Draw the useful insights from it

2. Introduction

Play store is an Android Market that serves as the official app store for certified devices running on the Android Operating system. Developed and Operated by Google, launched on 6th March, 2012.

Approximately 3.48 million apps are in the Play store. Play store apps have their own features such as Ratings, Reviews, Size and more. From the problem statement given, we should analyze the given database and should come up with the key factors that increased the number of users, long term usage etc., the objective of this

project is to deliver insights to understand customer demands better and thus help developers to popularize the product.

3. Steps involved:

The operations on the dataset are done by python scripts on Google Colab.

A. Importing Libraries:

Following Libraries are used in this analysis

i. NumPy: NumPy is a Python package. It stands for 'Numerical Python'. It is a library consisting of multidimensional array objects and a collection of routines for processing of array.

ii. Pandas:
Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with structured
and time series data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python.

iii. Matplotlib: Matplotlib is a visualization library in Python for 2D plots of arrays

iv. Seaborn: Seaborn is a library for making statistical graphics in Python

B. Data Cleaning:

1) Finding garbage values from all the columns and drop the rows

2) Drop the found garbage containing rows

3) Converting 'Size' column into valid numeric column

4) Converting 'Reviews' column into valid numeric column

5) Converting 'Installs' column into valid numeric column

C. Creating the visualizations:

Using the various functions of above-mentioned libraries, we have created various types of charts like Correlation heatmap, Pie chart, Bar Plot, Scatter Plot, Scatter Map Box, etc. to establish meaningful relationships between the variables of the Data set. You can see those charts in the attached code file.

4. Observation

1) In the free apps 64.1 % apps has positive sentiment followed by 22.4 % and 13.6 % for Negative and Neutral sentiment respectively

2) In the paid apps 78.1 % apps has positive sentiment followed by 15.1 % and 6.7 % for Negative and Neutral sentiment respectively

3) From the Sentiment , review graph we can see that no of installs 1000000 and 10000000 has highest review with positive sentiments

4) Installs and review has a positive correlation

5) total 92.2 % apps are free and 7.8 % apps are paid in playstore data

6) Family app category contains highest free and paid apps

7) with the content rating analysis 81.8 % apps are of Everyone content rating

8) In analysis of the top Five categories in terms of apps count. Family has highest count i.e. 1829 apps and lowest Medical 395

9) In analysis of the top Five categories in terms of Installs. Game has highest installs and lowest is social

10) In analysis of rating pattern most of the app has rating between 4-4.5

11) In analysis of app size pattern most of the app has size between 1-10

12) In analysis of android version then most of the apps are fo 4.0 and up

13) If we consider app update month then most of the app update in jul

14) In analysis of app update year wise then update of app increasing year on year basis . highest updated in 2018

15) If we analyze the data in terms of revenue then Family category app has highest revenue

5. Conclusion:

After completion of the Project we have learned data cleaning and how to filter the data. We learned to differentiate between which data is essential and which is not from the given dataset and how to visualize it using various visualization techniques.

- Out of total 92.2 % apps are free and in it 64.1 % apps has positive sentiment
- Apps has 1000000 , 5000000 and 10000000 has positive sentiment more
- Family category has most apps free as well as paid. And also family category app has highest revenue
- Most of the apps build which are for everyone content rating
- Reviews and installs are positively correlated
- Apps updates are increasing year on year basis